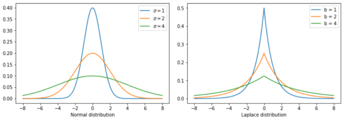
Introduction to Machine Learning

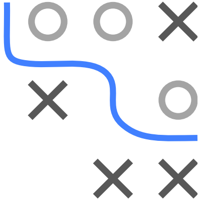
Regularization and Bayesian Priors



Learning goals

- Know how regularized risk minimization is same as MAP in Bayesian perspective
- Know correspondence of Gaussian/Laplace priors and $L2/L1$ regularization

RRM VS. BAYES



We already created a link between max. likelihood estimation and ERM.

Now we will generalize this for RRM.

Assume we have a parameterized distribution $p(y|\boldsymbol{\theta}, \mathbf{x})$ for our data and a prior $q(\boldsymbol{\theta})$ over our parameter space, all in the Bayesian framework.

From the Bayes theorem we know:

$$p(\boldsymbol{\theta}|\mathbf{x}, y) = \frac{p(y|\boldsymbol{\theta}, \mathbf{x})q(\boldsymbol{\theta})}{p(y|\mathbf{x})} \propto p(y|\boldsymbol{\theta}, \mathbf{x})q(\boldsymbol{\theta})$$

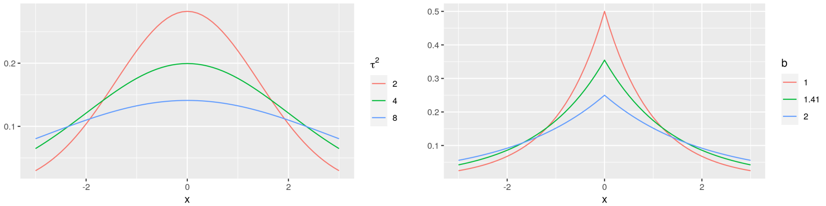
RRM VS. BAYES

The maximum a posteriori (MAP) estimator of $\boldsymbol{\theta}$ is now the minimizer of

$$-\log p\left(y \mid \boldsymbol{\theta}, \mathbf{x}\right) - \log q(\boldsymbol{\theta}).$$

- Again, we identify the loss $L\left(y, f(\mathbf{x} \mid \boldsymbol{\theta})\right)$ with $-\log(p(y|\boldsymbol{\theta}, \mathbf{x}))$.
- If $q(\boldsymbol{\theta})$ is constant (i.e., we used a uniform, non-informative prior), the second term is irrelevant and we arrive at ERM.
- If not, we can identify $J(\boldsymbol{\theta}) \propto -\log(q(\boldsymbol{\theta}))$, i.e., the log-prior corresponds to the regularizer, and the additional λ, which controls the strength of our penalty, usually influences the peakedness / inverse variance / strength of our prior.

RRM VS. BAYES



- *L*2 regularization corresponds to a zero-mean Gaussian prior with constant variance on our parameters: $\theta_j \sim \mathcal{N}(0, \tau^2)$
- *L*1 corresponds to a zero-mean Laplace prior: $\theta_j \sim \mathit{Laplace}(0, b)$. $\mathit{Laplace}(\mu, b)$ has density $\frac{1}{2b}\exp(-\frac{|\mu - x|}{b})$, with scale parameter b, mean μ and variance $2b^2$.
- In both cases, regularization strength increases as the variance of the prior decreases: a prior probability mass more narrowly concentrated around 0 encourages shrinkage.
- Elastic-net regularization corresponds to a compromise between Gaussian and Laplacian priors ▸ Zou and Hastie, 2005 ▸ Hans, 2011

EXAMPLE: BAYESIAN L2 REGULARIZATION

We can easily see the equivalence of *L*2 regularization and a Gaussian prior:

- We define a Gaussian prior with uncorrelated components for $\boldsymbol{\theta}$:

$$q(\boldsymbol{\theta}) = \mathcal{N}_d(\mathbf{0}, diag(\tau^2)) = \prod_{j=1}^{d} \mathcal{N}(0, \tau^2) = (2\pi\tau^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2\tau^2}\sum_{j=1}^{d}\theta_j^2\right).$$

- With this, the MAP estimator becomes

$$\begin{aligned}
\hat{\theta}^{\text{MAP}} &= \arg\min_{\boldsymbol{\theta}} \left(-\log p\left(y \mid \boldsymbol{\theta}, \mathbf{x}\right) - \log q(\boldsymbol{\theta})\right) \\
&= \arg\min_{\boldsymbol{\theta}} \left(-\log p\left(y \mid \boldsymbol{\theta}, \mathbf{x}\right) + \tfrac{d}{2}\log(2\pi\tau^2) + \frac{1}{2\tau^2}\sum_{j=1}^{d}\theta_j^2\right) \\
&= \arg\min_{\boldsymbol{\theta}} \left(-\log p\left(y \mid \boldsymbol{\theta}, \mathbf{x}\right) + \frac{1}{2\tau^2}\|\boldsymbol{\theta}\|_2^2\right).
\end{aligned}$$

- We see how the inverse variance (precision) $1/\tau^2$ controls shrinkage.